Representing composed meanings through temporal binding

Representing composed meanings through temporal binding

Hugh Rabagliati[1]

Leonidas A.A. Doumas[1]

Douglas K. Bemis[2]

[1]University of Edinburgh

and

[2]Geometric Intelligence, Inc.

Address for correspondence:

Hugh Rabagliati

School of Philosophy, Psychology and Language Sciences

University of Edinburgh

Edinburgh

EH8 9JZ

United Kingdom

hugh.rabagliati@ed.ac.uk

Abstract

A key feature of human thought and language is compositionality, the ability to bind pre-existing concepts and word meanings together in order to express new ideas. Here we ask how newly composed complex concepts are mentally represented and matched to the outside world, by testing whether it is harder to verify if a picture matches the meaning of a phrase, like *big pink tree,* than the meaning of a single word, like *tree*. Five sentence-picture verification experiments provide evidence that, in fact, the meaning of a phrase can often be checked just as fast as the meaning of one single word (and sometimes faster), indicating that the phrase's constituent concepts can be represented and checked in parallel. However, verification times were increased when matched phrases had more complex modification structures, indicating that it is costly to represent structural relations between constituent concepts. This pattern of data can be well-explained if concepts are composed together using two different mechanisms, binding by synchrony and binding by asynchrony, which have been suggested as solutions to the "binding problem" faced in both vision science and higher-level cognition. Our results suggest that they can also explain aspects of compositional language processing.

197 words.

Keywords

meaning, semantic composition, binding problem, prediction

Compositionality is a key feature of human thought and language: We can effortlessly combine older, more basic concepts and word meanings in order to express an unbounded number of new ideas. For instance, even though the words *Spotted, Pink,* and *Tree* are rarely juxtaposed, they can be quickly composed together to create a coherent semantic interpretation.

Work in linguistic semantics, philosophy, and psychology has considerably advanced our understanding of how complex concepts, such as the meanings of phrases, might be built from their component parts (Heim & Kratzer, 1998; Pylkkänen & McElree, 2006; Werning, Hinzen, & Machery, 2012). This has included discoveries about the role of broader world knowledge in interpreting the meanings of phrases (Barner & Snedeker, 2008; Springer & Murphy, 1992), and about the neural implementation of combinatorial operations (Bemis & Pylkkänen, 2011; Frankland & Greene, 2015; Pylkkänen & McElree, 2007).

However, amongst this research there is a surprising gap in our knowledge: we know little about how composed representations are held in mind in order to be matched against the world. While we know a great deal about how individual words (like *spotted, pink* or *tree*) are stored in working memory (Baddeley, 2003), and about how complex concepts can, with experience, be "chunked" into simple units (Cowan, Chen, & Rouder, 2004), we know much less about how newly encountered combinations of concepts are mentally represented. For example, how does the representation of a complex

concept, such as *big pink tree*, differ from the representation of a singleton concept, such as *tree,* or from the representation of a list of word meanings, such as *big, pink, tree*? Do complex representations, built by stacking ever more concepts, also demand ever more working memory? Can some complex concepts be stored in very efficient ways?

Some of the most relevant work has been on the idea of gist representations, the proposal that, as we read or listen to text, we discard our precise memories of the exact linguistic input and replace them with less precise summaries of that input's meaning. Theories of gist can explain how and why we discard less-relevant information about a sentence, but their accounts of meaning (in which, for example, sentences are recoded as sets of propositions Carpenter & Just, 1975; H. H. Clark & Chase, 1972; Kintsch, 1998) are more suited for explaining the representation of large chunks of text rather than characterizing the representations of simple concepts such as *big pink tree*. For example, it is unclear how the gists of *tree* and *pink tree* might differ. Potter (1993) has argued that gist representations are built by binding together token representations of concepts in a short term conceptual memory store. This idea seems plausible, but without a precise account of what these bindings might be like, it is hard to evaluate the implications of the claim for the questions posed at the start of this paper.

Potential insight into these bindings can be found in the literature on compositional connectionist models. In these models, individual concepts (i.e., word meanings) are stored as separate nodes in a large neural network. The concepts can be composed together (i.e., bound) through their simultaneous co-activation (so-called temporal binding). The key idea, which has its roots in the "binding by synchrony" hypothesis from visual attention (Singer & Gray, 1995; Von Der Malsburg, 1984), is that composed concepts like *pink tree* might be represented in a neural network by simultaneously activating nodes for the constituent concepts, i.e., *pink* and *tree* (Hummel & Holyoak, 1997). This idea has been implemented in a number of neural network models, such as Hummel and Holyoak's (1997) model of analogy formation. Importantly, it has also recently received support as a neurophysiologically plausible account of how combinatorial linguistic structure might be represented (Ding, Melloni, Tian, Zhang, & Poeppel, 2015).

One reason that binding by synchrony is a plausible candidate mechanism for compositional binding is that it provides an efficient way of compressing information, just like a gist. Because neural networks operate in parallel, the complexity of a network in which only *tree* is active is not importantly different from the complexity of a network in which both *pink* and *tree* are activated. That is to say, the network pays essentially no additional cost (e.g., at least in terms of storage) in order to represent *pink tree* as opposed to *tree.*

Binding by synchrony is therefore a clever and efficient default mode of representation. However, it displays an important difficulty accounting for certain types of more complex compositional representations. In particular, when using binding by synchrony it is not possible to represent the precise structure with which concepts should be bound (Doumas, Hummel, & Sandhofer, 2008). While the simultaneous activation of a set of concepts does indicate which of them should be bound together, it does not indicate which concepts should serve as arguments and which should serve as predicates. This makes it difficult to represent any sort of well-structured concepts. To illustrate, consider how synchrony might be used to represent the concepts *pink tree* and *dark pink tree*. *Pink tree* can be easily represented through synchrony: the simultaneous activation of *pink* and *tree* will activate the features associated with pinkness and with treeness, features that are best matched by a pink coloured tree. However, if we try to represent *dark pink tree* through synchrony, we will produce an extremely inaccurate representation. In this case, we would activate features associated with darkness, with pinkness, and with treeness. These features would be best matched by something that is simultaneously a dark tree (e.g., a tree in darkness), a dark pink colour, a pink tree, a tree with a dark pink colour, and so on. This is clearly not a typically intended meaning of *dark pink tree*.

To represent the structure of a composed concept in a neural network, it is necessary to somehow "screen off" individual component concepts from each other, to create the constituent relationships of the structure (e.g., ensuring

that *dark* modifies *pink* but not *tree*). This is not simple to do. One suggestion has been to use so-called conjunctive codes, in which each component concept is given a separate representation for each possible role that it might play (e.g., we would store two representations of *dark,* one for when it modifies another adjective [*dark pink*] and one for when it modifies a noun [*dark tree*]). But this solution has a theoretically unsatisfying consequence, as it assumes that every concept must have a different instantiation for each potential role that it might play. This sort of ambiguity of representation is inconsistent with a fundamental principle of compositionality, that the meaning of an expression should be a function of the meaning of its parts; in this case, the meaning of a part would be determined by its function in an expression.

An alternative approach, and the one that we focus on here, is that concepts may be screened off from each other by using binding via *asynchrony*, in which the pattern of activation of concepts over time distinguishes different thoughts and creates constituent structure. For example, a phrase like *dark pink tree* can be represented with a constituent structure of the form [[dark & pink] tree] by initially co-activating *dark* and *pink* (to indicate a dark pink colour), and then subsequently activating *tree* in isolation (to indicate that the bound concept “dark pink” should modify *tree*) (Doumas et al., 2008; Hummel & Holyoak, 2003). In this case, the initial period of activation would first activate features associated with *darkness* and *pinkness* (which would be well matched by a dark pink colour), and then features associated with treeness (which would be well matched by a *tree*). That is to say, given this pattern of

activation, *dark pink tree* would be well matched by a tree with a dark pink colour.[1]

The ideas of binding by synchrony and asynchrony suggest answers to the questions posed at the start of the paper about the nature of compositional representations. Because simple compositional concepts with minimal structure, such as *pink tree*, can be represented through simultaneous activation, then their representation does not importantly differ from the representation of a single word (i.e., it is the pattern of activity in a neural network at one single point in time). More complex and structured concepts, however, must be represented by activating the different components of a concept across time. That is to say, the system pays a cost for precisely representing structure (e.g., some models limit the number of timesteps available (Doumas et al, 2008); this provides an upper bound on working memory capacity).

A potential challenge

The ideas behind synchronous and asynchronous binding can easily map on to the processes involved in completing an experimental task such as sentence-picture verification, in which participants read a phrase and then verify if it matches a subsequent picture (H. H. Clark & Chase, 1972). If two concepts are bound through synchrony, such as *pink tree* in Figure 1 (left side), then the perceivable (e.g., visual) features associated with those

[1] And an alternative activation pattern could represent pink tree in darkness.

concepts will be activated in parallel. Each of these features can then be simultaneously checked against the input. This means that it should be as easy for participants to verify the meaning of a phrase (*pink tree*) as to verify the meaning of a single word (*tree*), assuming that the key features (here, colour and shape) can be extracted from the picture at similar speeds. When concepts are bound asynchronously (Figure 1, right side), then the predictions are different. For a phrase like *dark pink tree,* each key component must be activated at a different timepoint. First, *dark* and *pink* are co-activated, along with their visual features, and these are checked against the input in parallel. Meanwhile, *tree* is activated, and its features are checked against the input. This mixed parallel-serial process would cause participants to be slower to verify the meaning of asynchronously bound phrases than synchronously bound phrases.

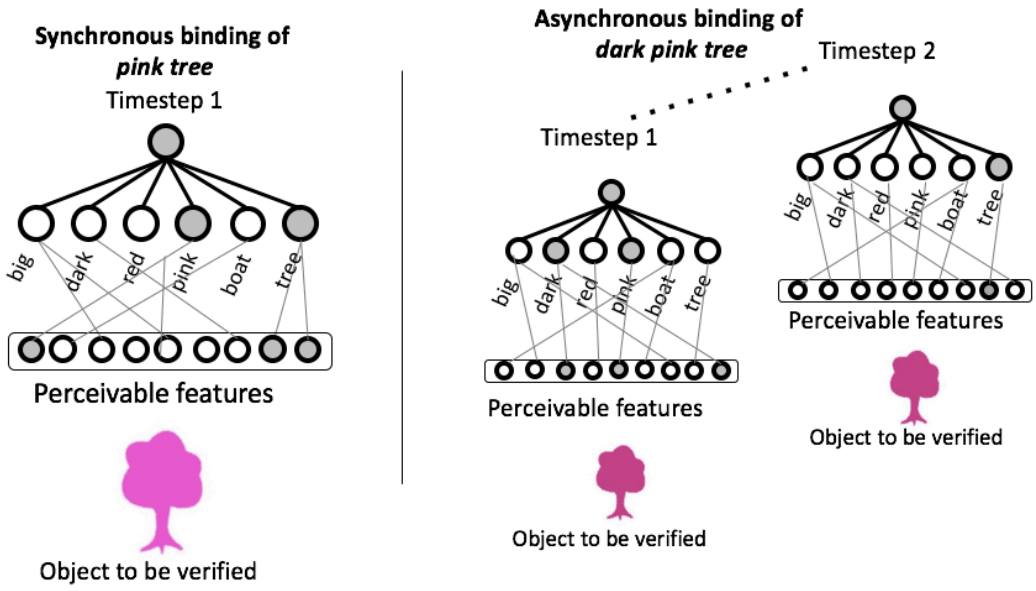

Figure 1. Illustration of synchronous and asynchronous binding, and how they relate to sentence-picture verification tasks.

However, one recent result suggests that both of these predictions – parallel checking and mixed parallel-serial checking – are incorrect, and that the meanings of some composed concepts can be verified faster than the meanings of single words. In a sentence-picture verification task (conducted as part of a magnetoencephalography study), Bemis and Pylkkanen (2011) found that participants were faster to verify whether pictures matched the meanings of previously-presented phrases (*pink tree*) than the meanings of previously-presented single words (*tree*). By contrast (and consistent with a long literature on verbal memory, Sternberg, 1969), participants were slower to verify that a picture matched one of two words presented in a list (*cup, tree*) than to verify if the picture matched a single word presented alone.

Although little-commented on at the time, this "composition advantage" is, intuitively, extremely surprising, as it seems so unlikely that a multiword phrase should be easier to represent and to check than a single word. It also runs entirely counter to the theories discussed above. If concepts are bound through synchrony then a composed concept should be no harder to verify than a single word. If a concept is bound through asynchrony then it should be harder to verify than a single word (as each additional constituent requires its own timestep). Under neither account should a phrase be easier to verify than a single word.

Given these considerations, the composition advantage, if robust, could provide important insights into the nature of compositional representations. To our knowledge, Bemis and Pylkkänen's result is the only clear demonstration of this effect, although Potter and Faulconer (1979) found a similar effect using more complex stimuli.

Here, we characterize both the nature of this compositional advantage and its consequences for models of concept composition. To preview, Experiments 1a-c confirm that composed, multiword expressions can – in certain situations – be checked faster than a single word. However, this advantage is driven by predictability – the extension of the phrase *pink tree* is smaller than *tree* – and when phrase and word are matched on predictability we find that they are processed with the same speed. Importantly, this suggests that the representation of a two-word phrase's meaning is not importantly different from the representation of a single word's meaning, which is consistent with the idea of synchronous binding.

Experiments 2 and 3, however, suggest that not all compositional concepts are represented in this fashion. In particular, we provide evidence that it is harder to perform sentence-picture verification for phrase meanings that cannot be accurately represented through synchronous binding (i.e., where one concept needs to be screened off from another), and which demand asynchronous binding instead. We therefore suggest that both synchronous and asynchronous binding are used to represent complex concepts.

Experiment 1

Experiment 1a attempted to replicate the composition advantage found in Bemis and Pylkkänen (2011). We subsequently tested two potential explanations for this advantage, one based on the particular syntactic characteristics of Adjective-Noun phrases (Experiment 1b) and one based on the increased semantic specificity of a phrase like *pink tree* compared to *tree* alone (Experiment 1c).

In all the experiments reported here, we used words whose meaning referenced easily-depicted visual features, in particular shape, size, color, and texture, which can be detected with minimal effort by participants. As such, the critical reaction time differences in our experiments are difficult to explain based on the visual stimuli and instead, we would argue, are a consequence of how combined concepts are stored and verified.

Experiment 1a

Participants

20 participants with American I.P. addresses recruited from Amazon Mechanical Turk and paid $2.10 for participation. Sample size was set based on our intuition about the effectiveness of the manipulation.

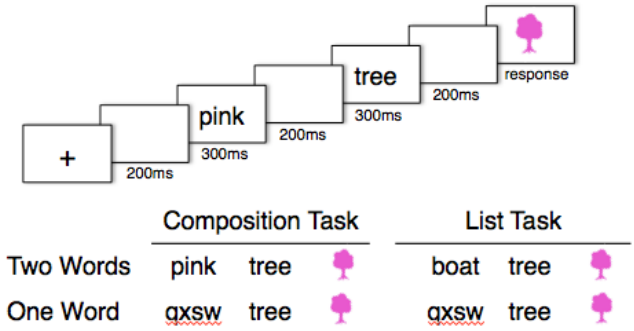

Figure 2. Procedure and conditions for Experiment 1a.

Materials and Procedure

Our procedure (Figure 2) followed Bemis and Pylkkänen (2011). In the composition task, participants read either single words (*tree*) or two word phrases presented word-by-word (*pink tree*). Single words were preceded by a nonsense string to equate the length of the trial. Participants then saw a coloured shape and decided whether it matched the preceding word(s). The same colored shapes were used in each condition, and were displayed at one of three random orientations to make the task more difficult. Shapes mismatched the word(s) on half of the trials. In two-word trials, pictures could mismatch through either colour or shape; in one-word trials pictures could only mismatch through shape.

In the list task, participants read either single words (*tree*, as in the composition condition) or a list of two nouns (*boat, tree*). Again, single words were preceded by a nonsense string. Participants then saw a coloured shape and indicated if it matched the preceding words. On two-word list trials,

participants simply had to decide if the shape matched either of the two previous words.

Each trial began with a fixation cross, followed by an initial word (or nonsense string) and then a second word. Words were displayed in 18-point font and presented on screen for 300ms followed by a 200ms inter-stimulus interval. Pictures were displayed until the participant responded.

In the composition task, trials were created by randomly combining one of twenty-five nouns [boat, house, lamp, star, disc, plane, bag, lock, cane, hand, key, shoe, bone, square, bell, bow, car, cross, cup, flag, fork, heart, leaf, note, tree] with one of six adjectives [red, blue, pink, black, green, brown] or with length-matched non-words [xkq, qxsw, mtpv, rjdnw, wvcnz, zbxlv][2]. In the one-word condition, the colour and orientation of the subsequent picture were randomly chosen. In the two-word condition, only the orientation of the picture was randomly chosen.

In the list task, trials were created by randomly combining one of the twenty-five nouns with either another noun or a non-word (which always preceded the noun). In both the one-word condition and the two-word condition, the colour and orientation of the subsequent picture were randomly chosen.

[2] We removed the non-words in a replication of Experiment 2, such that trials differed in length, but found the same pattern of results, suggesting their use does not affect the experimental outcome.

Participants viewed 100 trials in each of the four conditions (50 match and 50 mismatch). The order of the tasks (composition or list) was blocked between participants. The experiment was created using JSPsych (de Leeuw, 2014) and administered using PsiTurk (Gureckis et al., in press). Before beginning each block, participants completed 16 practice trials with feedback.

This research was approved by the Psychology Research Ethics Committee of the School of Philosophy, Psychology, and Language Sciences, University of Edinburgh. Data and analysis scripts for all reported experiments can be found at https://github.com/hughrabagliati/CompositionalityPaper1/.

Analysis and Results

We excluded trials with reaction times lower than 300ms or greater than 1500ms (criteria used throughout the paper). This removed 921 trials out of 7616 (12% of the total, median 17.5 per subject (SD=85, with the high standard deviation driven by two subjects)). We analyzed reaction times (on correctly answered match trials) using a linear mixed effects model (which accounts for imbalanced data) of the form (in lmer syntax) RT ~ Length * Task Type + (1+Length*Task Type |Subject) + (1+Length*Task Type|Item). This model predicts reaction time based on effects of Phrase Length (One or Two words), Task Type (Composition or List) and their interaction, and includes random intercepts for subjects and items, and the maximal random effects structure that permitted convergence. We calculated *p* values via model comparison. Since accuracy data were proportional, they were analyzed using

a similar mixed-effects logistic regression of the form (in glmer syntax) Accuracy ~ Length * Task Type + (1+Length +Task Type |Subject). In this regression, the by-items random effects structure, and the by-subjects interaction slope between Length and Task Type had to be dropped to aid convergence (there was little variability in either of these terms in the more maximal regression that did not converge). In all regressions reported in this paper, we use the maximal random effects structure that permits convergence.

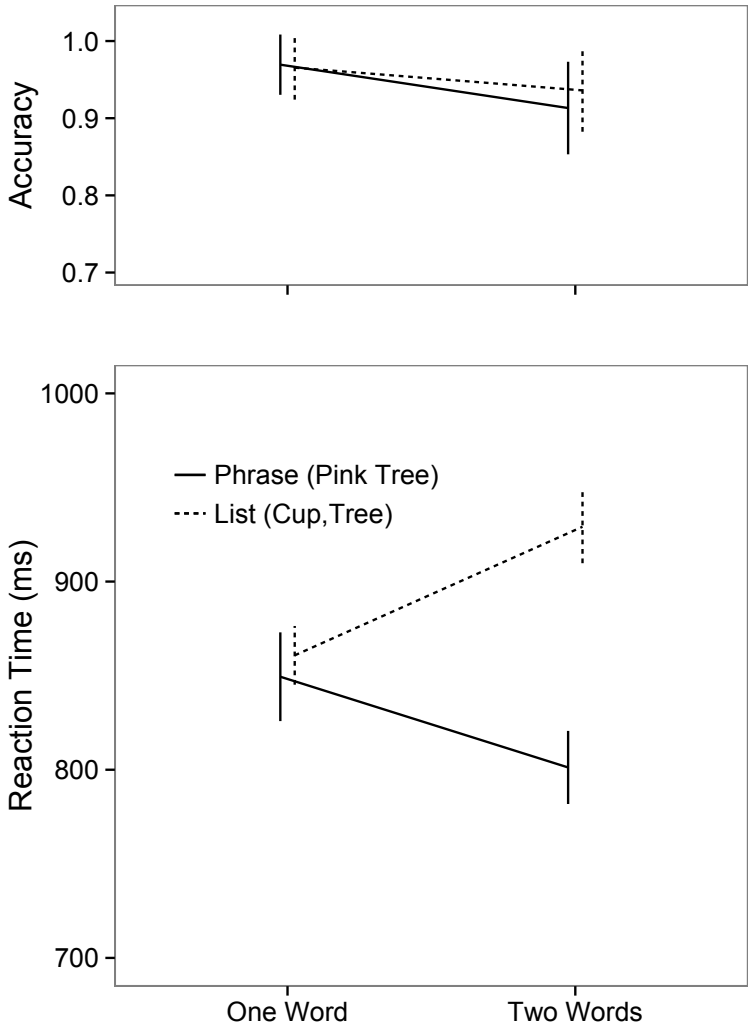

Figure 3. Top: Mean accuracy at identifying pictures in Experiment 1a. Bottom: Mean reaction time to correctly identify matching pictures in Experiment 1a. Error bars show +/- 1 within-subject standard error (all conditions are within subjects).

As seen in Figure 3, we replicated the critical reaction time result from Bemis and Pylkkänen (2011), an interaction between length and task type (β=25.0 (5.5), $\chi 2(1)$=13.6, p<.001): In the composition task, participants were faster to verify two-word phrases than single words (β=-19.3 (8.8), $\chi 2(1)$=4.4, p=.04), but in the list task they were slower to verify two-word lists than single word lists (β=31.7 (7.1), $\chi 2(1)$=13.1, p<.001).

Importantly, this interaction in reaction times was not well explained by a speed accuracy trade-off. In our accuracy analysis the interaction was not reliable (see Figure 3, top panel, β=0.06 (0.07), $\chi 2(1)$=0.78, p=.38). Experiment 1 therefore confirms Bemis and Pylkkänen's (2011) evidence for a "composition advantage" in checking complex concepts.

	One Word	Two Words
Phrase	849 [772,923]	801 [737,875]
List	860 [774,934]	929 [851,1009]

Table 1. Mean reaction times [and 95% confidence intervals] in milliseconds for Experiment 1a.

Experiment 1b

Experiment 1a clearly replicated the composition advantage: Participants verified composed phrases (*pink tree*) faster than either single words (*tree*) or lists of words (*boat, tree*). However, a potential objection to these comparisons is that *pink tree,* unlike the other stimuli, contains an adjective. That is to say, our participants were not only performing a different task between the list and the composition condition, but were performing that task on different stimuli. In Experiment 1b we tested whether the advantage still holds when phrases and lists are more precisely matched, by using adjective-noun pairs in both the composition and list conditions. For example, participants in the list condition might now have to verify whether a picture was pink or a tree.

Methods

Participants

12 participants with American I.P. addresses recruited from Amazon Mechanical Turk and paid $4.50 for participation. We had intended to test 20 participants, but stopped early after participants complained that the Adjective-Noun list task was too hard. We additionally excluded one participant who contacted us to say that they did not follow the instructions

(due to the difficulty), and one participant whose accuracy was 0 in the list task.

Materials and Procedure

We followed the same experimental parameters as Experiment 1a unless otherwise noted. The list condition was the same as the composition condition (e.g., adjectives preceded nouns) except that participants subsequently saw pictures that either matched only one of the words (e.g., reading *pink tree* and seeing a pink disc) or neither.

Results

Reaction times were submitted to a mixed effects model that predicted reaction times based on fixed effects of Phrase Length (One or Two words), Task Type (Composition or List), and their interaction, as well as the maximal random effects structure that permitted convergence: random intercepts for subjects and items, random by-subject slopes for phrase length and task, and a random by-item slope for phrase length. In lme4 syntax, this model had the form RT ~ Length * Task Type + (1+Length + Task Type|Subject) + (1+Length|Item). Accuracy was analyzed using a similar logistic regression with the form Accuracy ~ Length * Task Type + (1+Length|Subject) + (1+Length|Item), i.e., the random by-subject slope for task was removed to ease convergence. We excluded 161 trials out of 4800 (3% of the total, median 11 per subject (SD=12)) based on the criteria laid out in Experiment 1a.

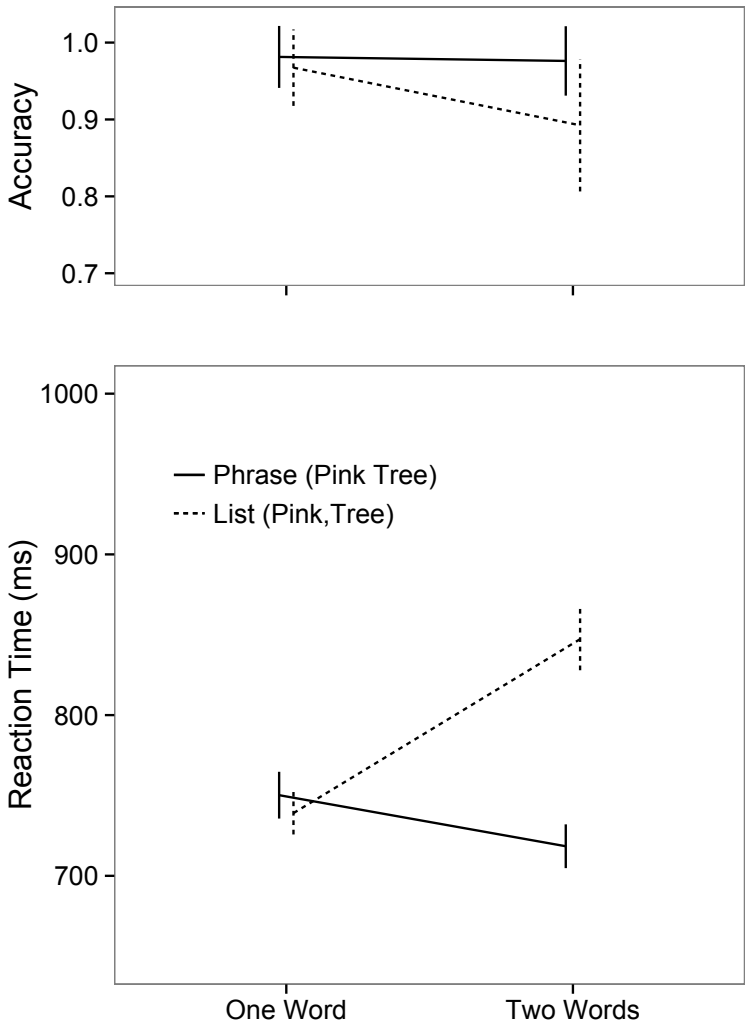

Figure 4. Top: Mean accuracy at identifying pictures in Experiment 1b. Bottom: Mean reaction time to correctly identify matching pictures in Experiment 1b. Error bars show +/- 1 within-subject standard error (all conditions are within subjects).

Our data, seen in Figure 4 and Table 2, replicate the findings of Experiment 1a while controlling for lexical features. We found a reliable two-way interaction between length and task (β=37.2(3.7), χ2(1)=97.7, p<.001): Participants were faster to verify two-word adjective-noun combinations than

single nouns in the composition condition (β=-16.2(5.0), $\chi2(1)$=8.1, p=.004) but they were slower to verify the exact same two-word stimuli when presented in the list task (β=58.6 (6.7), $\chi2(1)$=24.2, p<.001). Participants' accuracy was also greatly affected by this manipulation, indicated by a further Length by Task interaction (β=-0.25(0.09), $\chi2(1)$=7.3, p=.007): Compared to single words, accuracy was lower in the two word condition in the list task (β=-0.69(0.1), $\chi2(1)$=54.7, p<.001), but not the phrase task (β=-0.13(0.14), $\chi2(1)$=0.8, p=.37). [3]

These major effects on both reaction time and accuracy, found even though the words and phrases were held constant between the composition and list tasks, lead us to conclude that the composition advantage is not due to any special properties of adjectives, but is instead a property of composed representations compared to non-composed representations.

	One Word	Two Words
Phrase	750 [697,812]	718 [664,779]
List	739 [678,801]	847 [778,911]

Table 2. Mean reaction times [and 95% confidence intervals] in milliseconds for Experiment 1b.

[3] Note that reaction times in this experiment were somewhat faster than in Experiment 1a. It is not clear why reaction times were faster but, given that noise between-participants is high in online experiments, and given that our sample size for this study was quite small, we suspect that it is a statistical artifact, and indeed the difference was not statistically significant.

Experiment 1c

The set of things referred to by *pink tree,* called its extension, is smaller than the set of things referred to by *tree* (which includes trees of all colors). This difference could perhaps explain the composition advantage: Participants could more easily predict the correct picture for *pink tree* (pink trees at one of three orientations) than for *tree* (trees of six colours at one of three orientations). This explanation is given some prima facie plausibility by recent demonstrations that participants can rapidly translate linguistic information into predictions about the likely visual form of a referent (Rommers, Meyer, Praamstra, & Huettig, 2013; Zwaan, Stanfield, & Yaxley, 2002). However there is also an important reason for doubting it: there was no need for participants in our one-word condition to even attend to the color of the subsequent picture, and it is known that the shape of an object can be processed separately from its color (Garner & Felfoldy, 1970).

To test this predictability-based explanation of the composition advantage, in Experiment 1c we equated how easy it was for participants to predict what they would see after reading either a word or a phrase. In particular, we ensured that, on correct trials, participants could always predict the color of the picture that they would see.

Note that if this prediction-based explanation were to be correct, it would still leave open two possibilities for how composed phrases might be represented. Under binding by synchrony, *pink tree* and *tree* would take up the same amount of working memory resources and so should be verified with the same speed when their extensions are matched. Under binding by asynchrony, *pink tree* would take up more working memory than *tree* and should be verified more slowly when their extensions are matched.

Participants

24 participants with American I.P. addresses recruited from Amazon Mechanical Turk and paid $2.10 for participation. Our manipulation of predictability was between-subjects; we used the same size sample for each condition as Experiment 1b. We excluded one additional participant who failed to complete the task, and one who had very low accuracy.

Materials and Procedure

We followed the same experimental parameters as the composition condition from Experiment 1 unless otherwise noted. In the Mismatched Predictability condition, two-word phrases had smaller extensions than single words, as in Experiment 1a. In the Matched Predictability condition, phrases and single words had matched extensions: the same colour was always used for all matching pictures on single-word trials, with that colour varying across participants (e.g., one participant might only see red pictures on single-word trials).

Results

Reaction times were analyzed with a mixed effects model that predicted reaction time based on Phrase Length (One or Two words), Predictability (Mismatched or Matched), random by-subject and by-item intercepts, a random by-subject slope for phrase length, and random by-item slopes for phrase length and predictability.[4] In lme4 syntax, this had the form RT ~ Length * Predictability + (1+Length|Subject) + (1+Length + Predictability|Item). Accuracy was analyzed using a similar logistic regression of the form Accuracy ~ Length * Predictability + (1+Length|Subject) + (1+Predictability|Item), i.e., we dropped the by-item slope for phrase length to aid convergence. We excluded 329 trials out of 4800 (7% of the total, median 5.5 per subject (SD=19)) based on the criteria laid out in Experiment 1a.

[4] There was no by-subject slope for predictability, because this factor was between subjects.

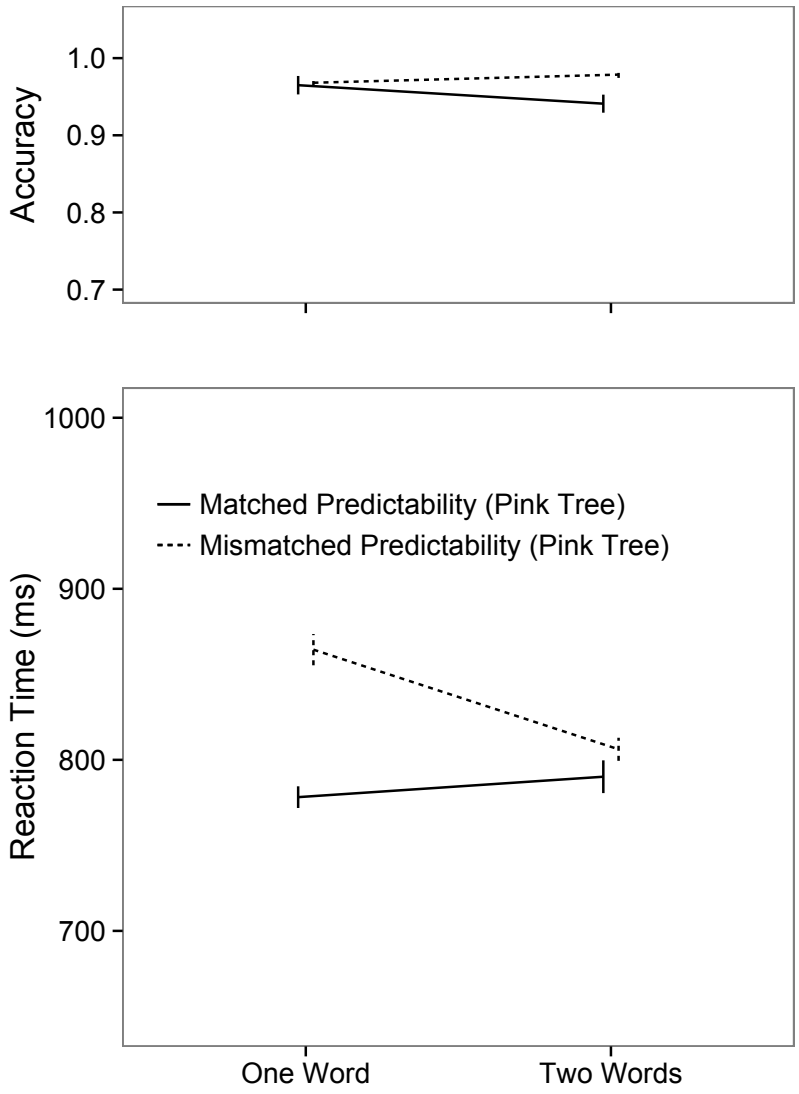

Figure 5. Top: Mean accuracy at identifying pictures in Experiment 1c. Bottom: Mean reaction time to correctly identify matching pictures in Experiment 1c. Error bars show +/- 1 within-subject standard error (Phrase Length is within subjects and Predictability is between subjects).

As seen in Figure 5 and Table 3, reaction times varied based on a length by predictability interaction (β=16.7(4.9), $\chi 2(1)$=11.7, $p<.001$). Just as in Experiments 1a and 1b, two-word phrases that were more predictive than single words were verified faster (β=28.8 (8.0), $\chi 2(1)$=10.2, $p=.001$). However,

when the predictability of phrases and single words were matched, they were checked with the same speed (β=-6.8(6.0), $\chi 2(1)$=1.2, p=.26), a finding that is notable because the predictability manipulation was implicit (i.e., participants were not told that there was only one possible match for single words).

We found no reliable effects on accuracy.

	One Word	Two Words
Mismatched Predictability	864 [758,973]	806 [709,906]
Matched Predictability	778 [692,853]	790 [711,860]

Table 3. Mean reaction times [and 95% confidence intervals] in milliseconds for Experiment 1c.

Experiment 2

Experiment 1c establishes two results. First, prediction strength can explain the composition advantage for multi-word phrases. Participants were as fast to verify that a picture matched a word as they were to verify that a picture matched a phrase, if they could predict the visual features of that picture with commensurable accuracy.

Second, and perhaps more important, Experiment 1c shows that, when matched for predictability, it takes participants the same amount of time to

verify the meaning of a two-word phrase as it does to verify the meaning of a single word. While this result is perhaps less surprising than the finding that phrases are processed faster than words, it is still important, and suggests that combinations of concepts can be mentally represented and checked in parallel, consistent with the idea that they are bound through synchronous activation (Hummel & Holyoak, 1997).

However, as discussed in the introduction, synchronous binding has difficulty representing concepts that have been composed together with any sort of complex structure. For example, if multiple concepts are activated simultaneously, then none of the concepts can be screened off from one another (i.e., all the concepts must compose directly). This means that synchronous binding cannot be used to accurately represent a concept such as *dark pink tree,* in which *dark* directly modifies *pink* but is screened off from *tree* (i.e., *dark pink* serves as a compound adjective). Under a model such as Doumas et al. (2008), representing this sort of structure would have to be done using asynchronous binding (e.g., *dark* and *pink* would need to be activated at one timestep, followed by *tree* at a second timestep [Figure 6]) meaning that the resulting representation would have twice the complexity.

One possibility is that we flexibly bring each binding scheme to bear depending on the representational demands of the task (Doumas et al., 2008). Asynchronous binding could be used when certain elements must be screened off, e.g., representing *dark pink tree* as *(dark pink) tree*.

Synchronous binding, which makes fewer demands on working memory, will be deployed otherwise (e.g., to represent *big pink tree*, in which the tree is both large in size and contains a large amount of pink).

If so, then participants should behave quite differently depending on the required coding scheme. Experiment 2 tests this. We examined how participants process single words (*tree*), two word phrases (*pink tree*) and three word phrases (*big pink tree*), all of which have a simple structure, and compared this to how participants process three word phrases that have a complex structure (*dark pink tree*), and so would need to be bound through asynchrony. Three word phrases with a simple structure that only require synchronous binding (*big pink tree*) should be checked as fast as two word phrases (*pink tree*). By contrast, three word phrases with a complex structure that demand asynchronous binding (*dark pink tree*) should take longer to retrieve than two word phrases.

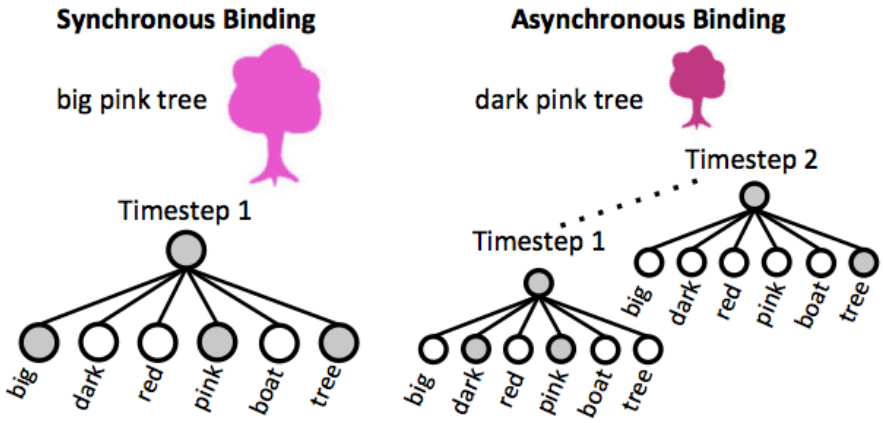

Figure 6. Illustration of synchronous binding (left) and asynchronous binding (right).

Methods

Participants

56 participants with American I.P. addresses recruited from Amazon Mechanical Turk and paid between $3.00 and $3.75 for participation. This sample size gave us approximately the same number of observations per condition as Experiment 1a. We excluded one additional participant who failed to complete the task.

Materials and Procedure

We followed the same experimental parameters as the composition condition of Experiment 1a unless otherwise noted. Participants read phrases of between one and three words and then judged if they matched a subsequent picture. Our binding type manipulation (Simple vs. Complex structure) was between subjects. Phrases were built by combining a noun with a colour adjective, as in Experiment 1, and then, for the simple structure condition, a size adjective [big/small] or, for the complex structure condition, a lightness adjective [dark/light]. Participants in the simple structure condition saw pictures that varied in shape, colour and size, while participants in the complex structure condition saw pictures that varied in shape, colour and saturation. Note that pictures in the simple condition were sometimes small (*small pink tree*) which may have made them slightly harder to identify. Participants completed 18 practice trials (with feedback) and 210 test trials.

Results

Reaction times were analyzed using a mixed effects regression that included fixed effects of Length (one, two, or three words), Structure (simple or complex) and their interaction, along with by-subject and by-item random intercepts, a by-subject random slope for Length, and by-item random slpes for Length and Structure. In lme4 syntax, this had the form RT ~ Length * Structure (simple or complex) + (1+Length|Subject) + (1+Length+Structure|Item). Length was coded as a numeric variable centred on zero (one word = -1, two words = 0, three words = 1). Accuracy was analyzed with a logistic regression with a roughly similar structure, but excluding by-item random effects because they did not converge (with the form Accuracy ~ Length * Structure + (1+Length+Structure|Subject)). Note that there was little variability in the Item term in the more maximal regression that did not converge. We excluded 635 trials out of 11760 (9% of the total, median 3 per subject (SD=20)) based on the criteria laid out in Experiment 1a.

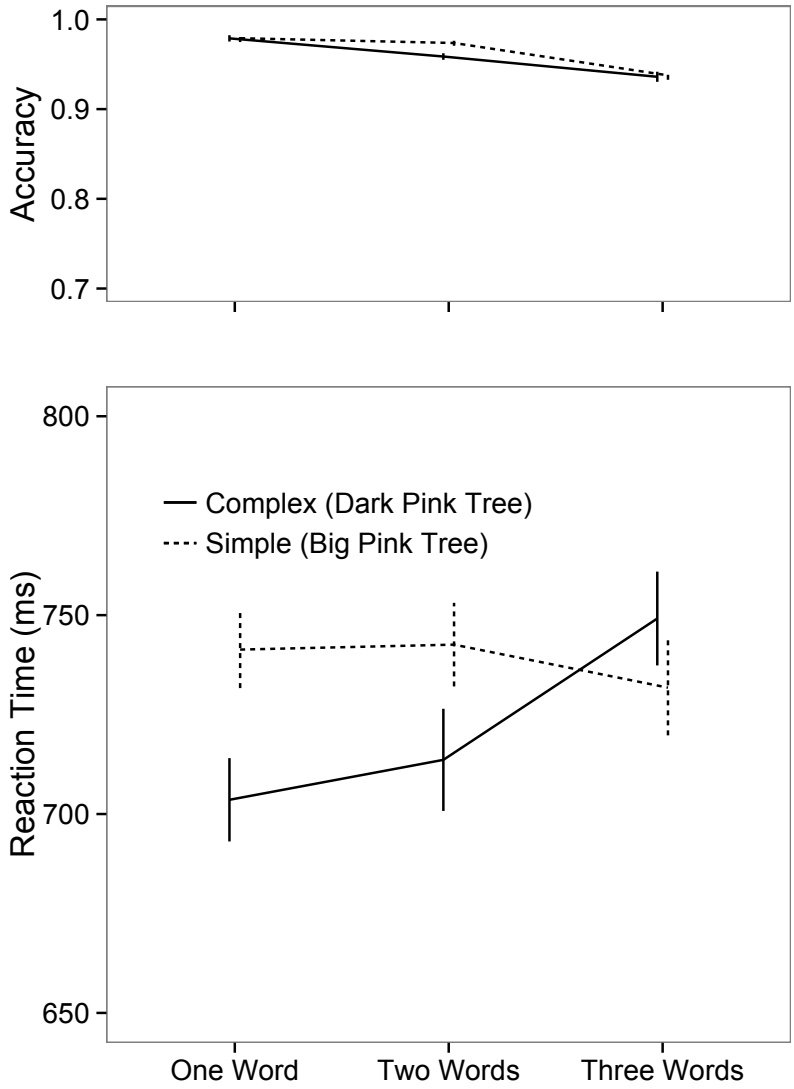

Figure 7. Top: Mean accuracy at identifying pictures in Experiment 2. Bottom: Mean reaction time to correctly identify matching pictures in Experiment 2. Error bars show +/- 1 within-subject standard error (Phrase Length is within subjects, and Structure is between subjects).

Consistent with our predictions, participants' reaction times displayed an importantly different profile across the two structure conditions, reflected in a Phrase Length by Structure type interaction (β=-13.4(5.0), $\chi 2(1)$=7.1, p=.007). In particular, as shown in Figure 7 and Table 4, we found a difference in how

three-word phrases were verified compared to two-word phrases across the two conditions (β=-22.3(9.0), χ2(1)=6.1, p=.01). In the simple structure condition (where we expected synchronous binding), we did not find any difference in the time participants took to verify three-word phrases (*big pink tree*) compared to two-word phrases (*pink tree,* β=-10.8(10.8), χ2(1)=1.0, p=.32). However in the complex structure condition (where we expected asynchronous binding), we found that participants took *longer* to verify three word phrases (*dark pink tree*) than to verify two-word phrases (*pink tree,* β=36.0(15.3), χ2(1)=5.2, p=.02).

Reaction times for one compared to two word phrases (neither of which needed asynchronous binding) did not interact with composition type (β=-4.6(8.1), χ2(1)=0.3, p=.57). However, there were two slightly surprising results here. First, we did not find a composition advantage in this study. Instead, one-word phrases were verified just as fast as two-word phrases (β=6.3(9.0), χ2(1)=0.5, p=.48) . We suggest that this is due to the decreased predictability of these more complex stimuli. Second, participants in the simple structure condition were numerically slower to process one and two word phrases compared to participants in the complex structure condition, although this difference was not statistically reliable (β=16.4(20.4), χ2(1)=0.66, p=.42). We suggest that this was due to difficulty identifying the smaller images on trials such as *small pink tree*. We return to both issues in the discussion section.

As in our other experiments, the critical reaction time differences cannot be explained by a speed-accuracy trade-off. Accuracy was unaffected by structure type and, critically, there was no interaction between phrase length and structure type: Participants were just as accurate at judging phrases of each length for both conditions (all $p > .25$). However, participants were overall less accurate when judging longer phrases (β=-0.6(0.08), $\chi 2(1)=44.8$, $p<.001$).

	One Word	Two Words	Three Words
Simple Structure	741 [677,816]	742 [680,812]	731 [672,795]
Complex Structure	703 [661,749]	713 [674,757]	749 [698,802]

Table 4. Mean reaction times [and 95% confidence intervals] in milliseconds for Experiment 2.

Experiment 3

Experiment 2's reaction time profiles suggest that participants were indeed using different types of representational format for the different composed concepts that we tested. The three word phrases that, we argued, did not require "screening off" and so could be represented using only synchronous binding (e.g., parallel activation of *big, pink* and *tree*), were checked as rapidly as single words or two word-phrases. In contrast, the phrases that, we argued, did require screening off and so required asynchronous binding (e.g.,

dark pink tree) were slower to be checked than either single words or two-word phrases, suggesting that the concepts could not be checked in parallel, which is consistent with asynchronous binding. This result therefore lends plausibility to our proposals about how combined concepts are held in mind.

Nevertheless, a potential confound in Experiment 2 is that the key adjectives in the critical three-word phrases were different: *dark* and *light* are less frequent than *big* and *small*, which may have led to different retrieval times in the three-word conditions.[5] We do not think that this confound explains our results: Although word frequency often affects reaction times, this dissipates with repeated presentation, and the key adjectives (*dark, light, big* and *small*) were repeated thirty-five times each. Still, it was a confound that we tried to overcome in Experiment 3.

In this experiment, participants saw identical one- to three-word phrases, but under conditions designed to push them to interpret the three-word phrases using either complex or simple modification structures. Phrases were always of the form: *big spotted tree, small striped disc,* etc (see Figure 8). In the Complex Adjective condition, participants were told that adjectives like *big* should modify the noun: phrases like *big spotted tree* were followed by pictures of big trees, but with spots that could be either big or small (i.e., participants had to screen off *big* from *spotted*). In the Complex Compound

[5] The *dark/light* condition also used a different set of pictures, but this cannot explain the result, as we did not find any effects in the one- and two-word conditions.

Adjective condition, participants were told that adjectives like *big* should modify adjectives like *spotted: big spotted tree* was followed by trees that had big spots but which were themselves of variable size (participants had to screen off *big* from *tree*). However, in the Simple condition, *big spotted tree* was followed by pictures of big trees with big spots, i.e., screening off was unnecessary, and participants could use synchronous binding. We predicted that, in the simple condition alone, three word phrases would be as easy to verify as two word phrases, while otherwise three word phrases should be harder to verify.

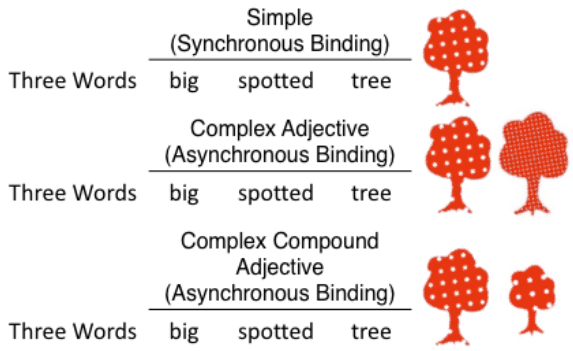

Figure 8. Examples of the three word conditions in Experiment 3.

Methods

Participants

60 participants with American I.P. addresses recruited from Amazon Mechanical Turk and paid $3.75 for participation. This sample size gave us approximately the same number of observations per condition as Experiment 1a. The 20 participants in the Complex Compound Adjective condition were

tested approximately two months later, inspired by comments about the work in a seminar. We excluded two participants who failed to complete the task.

Materials and Procedure

We followed the same experimental parameters as Experiment 2 unless otherwise noted. Phrases were built by combining a size adjective [*big/small*], followed by a texture adjective [*spotted/striped*] followed by one of twenty-three nouns [*tree, house*, etc]. Participants then saw a picture that varied in size, texture, shape, orientation, and colour (either red or blue).

Participants in the complex adjective condition were instructed that adjectives like *big* should modify the noun: they were told that if they read the phrase *big striped tree* and then saw a picture of a small tree with stripes, then they should reject the image, as the tree was small, not big. They then saw pictures in which the size of an object's stripes or spots was independent of its overall size (e.g., big spotted trees could have big or small spots). Participants in the complex compound adjective condition were instructed that adjectives like *big* should be interpreted as modifying the subsequent adjective: they were told that if they read the phrase *big striped tree* and then saw a picture of a tree with small stripes, then they should reject the image, as the tree's stripes were small, not big. They then saw pictures in which the overall size of the image was independent of the size of the image's spots or stripes. Participants in the simple structure condition were not given precise instructions as to what *big* should modify, but always saw pictures in which

the object's size and the size of its stripes/spots matched. As in Experiment 2, Structure type (simple/complex adjective/complex compound adjective) was varied between participants.

Participants completed 18 practice trials (with feedback) and 300 test trials, with a break halfway through.

Results

Reaction times were analyzed used a mixed effects model containing fixed effects of Phrase Length (one, two or three words), Composition type (C Complex Adjective, Complex Compound Adjective or Simple), their interaction, by-subject and by-item random intercepts, by-subject random slopes for Length, and by Item random slopes for Length. This had the form, in lme4 syntax, RT ~ Phrase Length * Composition type + (1+Length|Subject) + (1+Length|Item). Length was again coded as a centred numeric variable. The factor Composition type had three-level (that were sum coded); its main effect was thus represented by two predictors in the regression, as was its interaction with Length. We report Beta scores for both of the resulting predictors, but only one *p* value, as those were obtained via model comparison. Accuracy was assessed using a similar logistic regression, but excluding random by-subject and by-item slopes for Length to aid convergence (with the form Accuracy ~ Length * Composition type + (1|Subject) + (1|Item)). We excluded 1611 out of 18000 trials (9% of the total,

median 15 per subject (SD=34)) based on the criteria laid out in Experiment 1a.

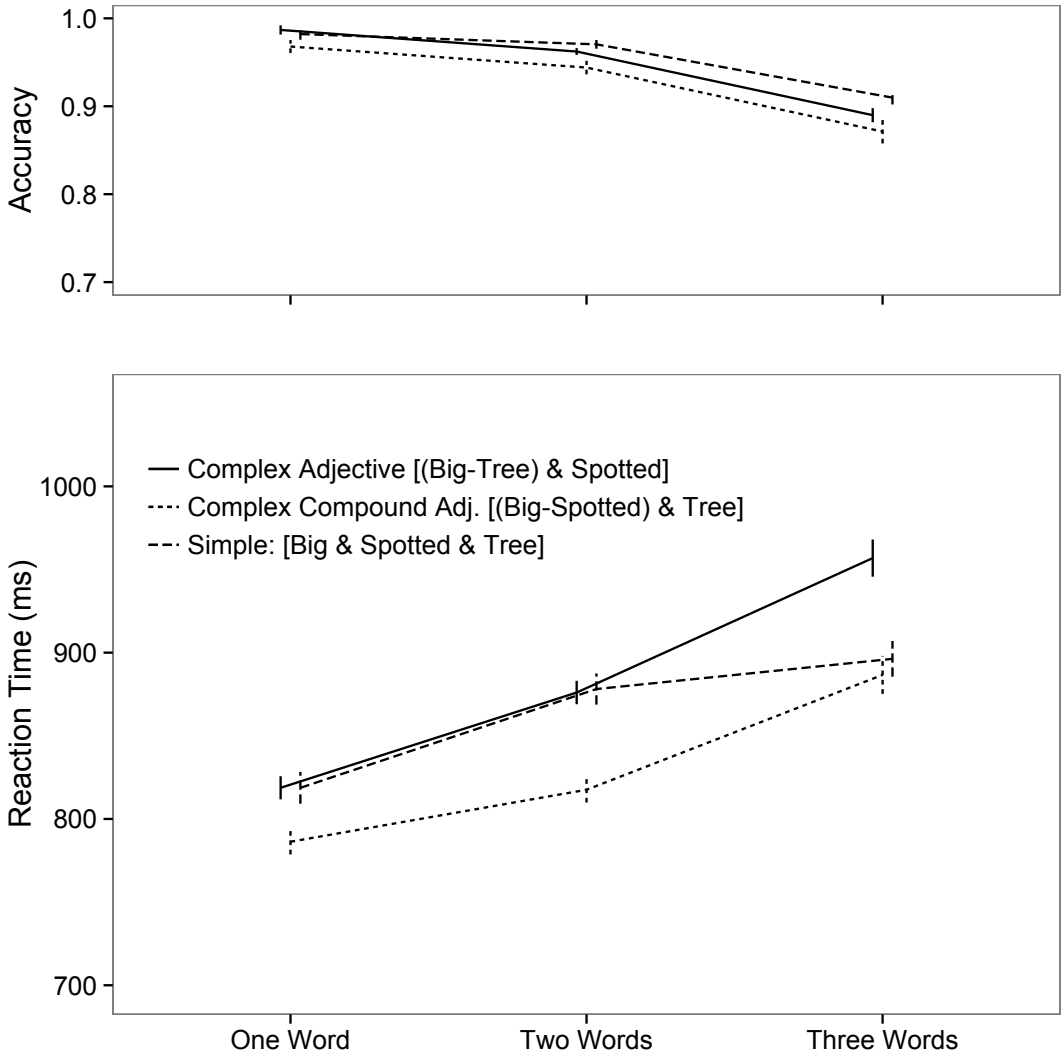

Figure 9. Top: Mean accuracy at identifying pictures in Experiment 3. Bottom: Mean reaction time to correctly identify matching pictures in Experiment 3. Error bars show +/- 1 within-subject standard error (Phrase Length is within subjects, and Structure is between subjects).

Just as in Experiment 2, participants' reaction times displayed an importantly different profile across the composition types, reflected in a Phrase Length by Composition type interaction (β_1=15.8(5.7), β_2=-12.7 (5.7), $\chi2(2)$=11.5, *p*=.003). As shown in Figure 9 and Table 5, we found important differences between how two- and three-word phrases were verified across the three conditions, reflected in a further Phrase Length by Composition type interaction when analysis was restricted to this subset (β_1=24.8 (11.1), β_2=-38.2(11.1), $\chi2(2)$=7.3, *p*=.015). In the Simple condition, where participants were expected to use synchronous binding, we did not find any difference in the time taken to verify three-word phrases compared to two-word phrases (β=18.6(14.0), $\chi2(1)$=1.8, p=.18). However, in both the Complex Adjective condition and the Complex Compound Adjective condition, we found that participants took *longer* to verify three word phrases than to verify two-word phrases (Complex Adjective: β=81.7(13.1), $\chi2(1)$=21.8, p<.001; Complex Compound Adjective: β=71.28(13.6), $\chi2(1)$=17.8, p<.001), as was expected if asynchronous binding was used in these conditions.

Importantly, the Composition manipulation had no further effects. When our analysis was restricted to the two- and one-word conditions, there was no reliable interaction between Composition type and Phrase Length (although the effect was marginal β_1=7.7 (8.4), β_2=12.0(8.5), $\chi2(1)$=5.5, p=.06) and no overall effect of Composition type (β_1=15.3(29.5), β_2=14.3(29.5), $\chi2(2)$=1.1, *p*=.59), although as can be seen in Figure 9, the Complex Compound Adjective participants were numerically faster. These null effects are

important, as they show that the critical three-word effect is unlikely to be explained by the small difference in predictability of our pictures, as predictability would have also caused a strong effect for shorter phrases.

However, and in contrast to all our previous results, two-word phrases took longer to verify than one-word phrases in this experiment (β=49.9(6.2), χ2(1)=44.3, p<.001). This difference may be due to the increased length, frequency or morphological complexity of our two texture adjectives (*spotted* and *striped*) compared to the colour adjectives used in previous experiments (e.g., *pink*), or to participants requiring a longer amount of time to verify the visual texture of the picture (there is some evidence that shape and texture are not "separable" features, and cannot be processed in parallel in the same way as shape and color (Garner & Felfoldy, 1970; Kimchi & Palmer, 1985))

As in our other experiments, the critical reaction time differences could not be explained by a speed-accuracy trade-off. There was no interaction between Phrase Length and Composition type: Participants were just as accurate at judging phrases of each length across the composition conditions (β_1=-0.12(0.08), β_2=-0.00008(0.08), χ2(2)=3.93, *p*=.14). As in Experiment 2, participants were less accurate when judging longer phrases (β=-1.02(0.05), χ2(1)=468, p<.001).

	One Word	Two Words	Three Words
Simple Structure	818 [744,907]	878 [808,961]	896 [830,976]
Complex Adjective	818 [751,888]	876 [810,944]	956 [887,1021]
Complex Compound Adjective	786 [722,845]	816 [747,883]	886 [811,957]

Table 5. Mean reaction times [and 95% confidence intervals] in milliseconds for Experiment 3.

General Discussion

Compositionality allows us to create new ideas by combining basic thoughts. Here, we investigated how composed concepts, and the structural relationships between them, are held in mind. We used a series of sentence-picture verification experiments, in which participants judged whether pictures matched noun phrases, to test how the complexity of a phrase's meaning (its number of words, and the relations between those words) affects how it is held in memory and checked against the world. Experiments 1 through 3 found that when the meaning of a phrase had a simple structure, participants were as fast to verify that phrase's meaning as to verify the meaning of a word (and were sometimes faster). Experiments 2 and 3 demonstrated that adding additional complexity to the structure of a phrase also added an additional burden: participants showed increased reaction times when they needed to check meanings with complex structures.

These results are consistent with the idea that we use two different mental schemes to link ideas together. In one scheme, binding by synchrony, composed concepts are activated and checked in parallel. This simultaneous activation binds concepts together without regard to structure. In the second scheme, binding by asynchrony, concepts are activated and checked in a mixed parallel-serial fashion, which maintains important structural relations between concepts (e.g., allowing the creation of constituent structure) at a cost to working memory. These schemes are both naturally implemented in compositional connectionist models such as DORA (Doumas et al., 2008) and LISA (Hummel & Holyoak, 2003).

However, while these models do a good job of accounting for the key phenomena uncovered here, they cannot account for all of the data presented. For example, they do not naturally explain the "composition advantage" found in both Bemis & Pylkkänen (2011) and also our own Experiment 1: Participants were faster to verify the meanings of two word phrases than single words when not matched on predictability. Indeed, these models predict only that participants should be as fast to verify a two-word phrase as a single word. To explain this discrepancy, Experiment 1c suggests that the composition advantage is due to predictive processes that are outside the purview of models such as DORA and LISA. In particular, participants appear able to rapidly transform entirely novel compositional concepts into accurate predictions about visual stimuli (see also Rommers et al., 2013;

Zwaan et al., 2002): When predictive strength is matched between one-word and two-word stimuli, participants perform in a manner consistent with the predictions of models such as DORA and LISA.

Our data leave open how these predictions might be generated; but our findings dovetail with other recent work suggesting that the meanings of words can be rapidly translated into predictions about related visual features. For example, participants are more likely to see a picture that is masked using continuous flash suppression (a form of binocular rivalry) when they can hear its name (Lupyan & Ward, 2013); since continuous flash suppression impedes access to the semantics of a masked image (Moors, Boelens, van Overwalle, & Wagemans, 2016), this result suggests that linguistic meanings are translated into low level visual features (see also Ostarek & Huettig, in press), a finding that aligns with predictive coding accounts of cognition, in which high-level knowledge is constantly used to generate lower-level predictions about the world, to facilitate perception and interaction (A. Clark, 2013).

Experiment 1's results thus indicate that compositional processes are quite powerful in how they quickly facilitate interactions with the world. Participants seem able to generate accurate predictions about both the colour and shape of named referents, and can check these predictions quickly. Future work should be concerned with the factors that might limit predictive compositional processes, in particular the degree to which multiple types of visual feature can be predicted simultaneously.

That said, one unexpected finding from Experiments 2 and 3 was the absence of a composition advantage. While the participants in Experiment 1a-c were repeatedly faster to verify two-word phrases than single words, the same was not true once three-word phrases were added into the study design in Experiments 2 and 3. One possibility, suggested in the results section, is that the increase in complexity of the visual stimuli (e.g., the addition of variation in size or shade) reduced participants' reliance on prediction. This could be because prediction was harder in the three-word condition (reducing participants' overall reliance on this strategy), because participants found it harder to generate accurate predictions in the one-word and two-word conditions (e.g., due to Experiment 2's unspecified variability in shape or shade), or for both reasons. Relatedly, in Experiment 3 we were surprised to find that participants were in fact slower to verify two word phrases than single words. As suggested in the results section, this could potentially be caused by slow processing of texture words in phrases like *spotted tree* or, perhaps, slow processing of visual texture information (e.g., participants may have to partially process both shape and colour before they are able to classify the picture's texture). A corollary of this is that multiple different factors, beyond prediction and binding type, are likely to affect verification times, including factors that may be specific to the type of stimulus being verified (e.g., the color of a stimulus may be easier to verify than its texture).

Beyond these less-expected data points, the ideas of binding by synchrony and asynchony made a number of interesting and unique predictions about verification times for phrases with similar surface features but importantly different internal complexity. These predictions were typically met. Still, because this work only examined the representation of simple phrases (and simple noun phrases at that) it leaves open a number of questions. In particular, it is important to study whether the synchronous and asynchronous binding schemes are sufficient to represent more complex modification structures, such as those involving long distance dependencies, that are often found in natural language (Martin, 2016). In addition, it would be interesting to understand how a system like this implements context-dependent meanings. In the current experiments we have assumed that words like *big, red* and *dark* make a constant contribution to the meaning of an expression, but in actuality the meanings of these words depend on the nouns that they modify (a *big mouse* is big for a mouse, a *red fox* is red for a fox, see Partee, 1995 for an introduction). One possibility is that words like *big* and *red* always make a relative contribution to the meaning of a phrase (i.e., *big* does not specify a particular size, but only specifies that the object it modifies is larger than its typical size); this would be easily implemented in a model such as DORA (Doumas et al, 2008).

Another important question concerns how the language processing system decides whether to use synchronous or asynchronous binding to represent the meaning of a phrase. The answer to this question is not currently clear.

One possibility is that the system has a default mode of representation, but it switches to the second mode when the task demands it. Synchronous binding is a good candidate for this default: Representing a meaning through simultaneous activation of all its components will be less costly, in terms of both energy and time, than representing a meaning through asynchronous activation of its components over time. The system may then automatically switch to using asynchronous binding when the task itself demands that more complex meanings be represented (e.g., when the reader realizes that they need to screen off *dark* from *boat*). In our experiments, participants may have realized that they needed to use asynchronous binding during the practice trials that preceded each task.[6] Future work will need to explain this mechanism more precisely.

The final issue that we will discuss – a critical one – concerns whether representational schemes other than synchronous/asynchronous binding might be able to explain the results reported here. Our experiments were designed to test specific ideas about how temporal binding may be used to create structured or structure-less representations of meaning, rather than being designed to rule out other formalisms, and so leave open the possibility that other schemes could also explain our key behavioral findings. For example, Natural Language Processing algorithms have recently achieved some considerable success parsing and representing the meanings of sentences by using Long Short Term Memory networks (LSTM networks,

[6] Trial order in the test trial blocks did not further interact with our core interactions

Hochreiter & Schmidhuber, 1997), which are variations on simple recurrent networks (Elman, 1990). LSTM networks are strikingly good at learning the statistics of their input, particularly at learning statistics over non-adjacent elements (e.g., allowing them to accurately model the dependencies in a sentence). If an LSTM network were to be trained on our task (using a sequence of words to predict a picture, and measuring prediction error as a proxy for reaction time), might the results be analogous to the behavior of our participants? While this question is beyond the scope of the present paper, we suspect that the answer could well be no. In particular, it is not clear how an LSTM network should (or even could) differentially represent a complex structure like [(Big-Spotted) & Tree] from a flat structure like [Big & Spotted & Tree], yet it is this representational distinction that drove the results of Experiments 2 and 3, above-and-beyond any effects of predictability. Indeed, there is some evidence that LSTM networks trained on unstructured input, as well as other approaches to semantic composition based on distributed representations of word meaning (Mitchell & Lapata, 2010), have difficulty representing certain basic linguistic relationships, such as scope or the similarity of meanings across sentences, which appear to demand more complex structured semantic representations (Bowman et al., 2016; Gershman & Tenenbaum, 2015). An important future goal should therefore be to develop architectures that can take advantage of both structured representations and also the statistical learning capabilities of LSTM networks; the current paradigm could potentially serve as a benchmark for evaluating this kind of model.

A second potential way of modeling our findings is in terms of the procedures used for verifying whether a meaning is matched in the world. Under the account that we tested, participants' verification times were affected by how they represent the meanings of phrases (synchronously versus asynchronously), and these representations affected how verification proceeded (e.g., when concepts were asynchronously bound, access to the component concepts took longer, and so verification took longer). But verification processes could also potentially affect reaction times under models that do not use synchronous versus asynchronous binding. For example, consider a model that is able to represent the two meanings of the ambiguous phrase *dark pink tree* (i.e., [(Dark-Pink) & Tree] and [Dark & (Pink-Tree)]). This could be done by combining a set of functions in accord with the phrase's syntactic structure, such as is done in modern formal semantic theories (although it is unclear how such trees themselves would be represented). If this system is able to represent both of the structured meanings, but can only check one meaning at a time, then it may be able to explain some of our results; in particular, the system should be slower to recognize a *dark pink tree* than a *pink tree*. This is because *pink tree* is unambiguous, and so only one representation can be verified, but *dark pink tree* is ambiguous, and so the system will have to choose which of the meanings should be verified first. If it picks the wrong representation (e.g., [Dark & (Pink-Tree)] in our experiments) then verification should be a slower process. This verification-based theory is an interesting possibility, but we do

not think that it is a particularly compelling explanation of our results, because we do not see a rational reason why participants should represent both meanings of the ambiguous phrases during our task (e.g., recall that participants never saw pink trees in darkness in Experiment 2). But it is still an interesting alternative interpretation, that deserves further investigation.

In concluding, we suggest that these findings provide insights that are distinct and complementary to a surge of recent interest in compositionality (Werning et al., 2012), including studies on the neural bases of compositional processes (Bemis & Pylkkänen, 2011; Ding et al., 2015; Frankland & Greene, 2015), and Bayesian models that approach the problem using Marr's so-called computational level of analysis (Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Piantadosi, Goodman, Ellis, & Tenenbaum, 2008). Our findings can help bridge these perspectives: The binding schemes suggested by this data reside at Marr's algorithmic level, i.e., mechanisms for implementing computational analyses in a way that might be neurally achievable. These mechanisms are plausible: binding through time has its roots in the attention and working memory literature (Singer & Gray, 1995) and, while controversial, has resulted in rich progress (Treisman, 1996). While the claim that these mechanisms are used in semantic composition clearly requires further investigation, our experiments so far suggest that they may have similar explanatory value for language and high-level cognition.

Word count:

Total: 9,191

Acknowledgments

This work was funded by an ESRC Future Research Leaders award (No. ES/L01064X/1) and by a Leverhulme Trust Research Project Grant (No. RPG-2014-253) to HR. Note that all data and analysis scripts can be found at https://github.com/hughrabagliati/CompositionalityPaper1.

References

Barner, D., & Snedeker, J. (2008). Compositionality and Statistics in Adjective Acquisition: 4-Year-Olds Interpret Tall and Short Based on the Size Distributions of Novel Noun Referents. *Child development, 79*(3), 594-608.

Bemis, D. K., & Pylkkänen, L. (2011). Simple composition: A magnetoencephalography investigation into the comprehension of minimal linguistic phrases. *The Journal of Neuroscience, 31*(8), 2801-2814.

Bowman, S. R., Gauthier, J., Rastogi, A., Gupta, R., Manning, C. D., & Potts, C. (2016). A fast unified model for parsing and sentence understanding. *arXiv preprint arXiv:1603.06021*.

Carpenter, P. A., & Just, M. A. (1975). Sentence comprehension: A psycholinguistic processing model of verification. *Psychological review, 82*(1), 45.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(03), 181-204. doi:doi:10.1017/S0140525X12000477

Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive psychology, 3*(3), 472-517.

Cowan, N., Chen, Z., & Rouder, J. N. (2004). Constant Capacity in an Immediate Serial-Recall Task: A Logical Sequel to Miller (1956). *Psychological Science, 15*(9), 634-640.

de Leeuw, J. R. (2014). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior research methods, 47*(1), 1-12.

Ding, N., Melloni, L., Tian, X., Zhang, H., & Poeppel, D. (2015). Cortical entrainment reflects hierarchical structure building in speech comprehension. *Nature neuroscience*.

Doumas, L. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological review, 115*(1), 1.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*(2), 179-211.

Frankland, S. M., & Greene, J. D. (2015). An architecture for encoding sentence meaning in left mid-superior temporal cortex. *Proceedings of the National Academy of Sciences, 112*(37), 11732-11737.

Garner, W. R., & Felfoldy, G. L. (1970). Integrality of stimulus dimensions in various types of information processing. *Cognitive psychology, 1*(3), 225-241.

Gershman, S. J., & Tenenbaum, J. B. (2015). *Phrase similarity in humans and machines.* Paper presented at the Proceedings of the 37th Annual Conference of the Cognitive Science Society.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A Rational Analysis of Rule‐Based Concept Learning. *Cognitive Science, 32*(1), 108-154.

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., . . . Chan, P. (in press). psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavioral Research Methods*.

Heim, I., & Kratzer, A. (1998). *Semantics in generative grammar* (Vol. 13). Oxford, UK: Blackwell.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation, 9*(8), 1735-1780.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: a theory of analogical access and mapping. *Psychological review, 104*(3), 427.

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological review, 110*(2), 220.

Kimchi, R., & Palmer, S. E. (1985). Separability and integrality of global and local levels of hierarchical patterns. *Journal of Experimental Psychology: Human Perception and Performance, 11*(6), 673.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*: Cambridge university press.

Lupyan, G., & Ward, E. J. (2013). Language can boost otherwise unseen objects into visual awareness. *Proceedings of the National Academy of Sciences, 110*(35), 14196-14201.

Martin, A. E. (2016). Language processing as cue integration: Grounding the psychology of language in perception and neurophysiology. *Frontiers in Psychology, 7*.

Mitchell, J., & Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive Science, 34*(8), 1388-1429.

Moors, P., Boelens, D., van Overwalle, J., & Wagemans, J. (2016). Scene Integration Without Awareness No Conclusive Evidence for Processing Scene Congruency During Continuous Flash Suppression. *Psychological Science*, 0956797616642525.

Ostarek, M., & Huettig, F. (in press). Spoken words can make the invisible visible–Testing the involvement of low-level visual representations in spoken word processing. *Journal of Experimental Psychology: Human Perception and Performance*.

Partee, B. (1995). Lexical semantics and compositionality. In L. Gleitman, M. Lieberman, & D. N. Osherson (Eds.), *An invitation to cognitive science: Language* (pp. 311-360).

Piantadosi, S., Goodman, N., Ellis, B., & Tenenbaum, J. (2008). *A Bayesian model of the acquisition of compositional semantics.* Paper presented at the Proceedings of the thirtieth annual conference of the cognitive science society.

Potter, M. C. (1993). Very short-term conceptual memory. *Memory & cognition, 21*(2), 156-161.

Potter, M. C., & Faulconer, B. A. (1979). Understanding noun phrases. *Journal of Verbal Learning and Verbal Behavior, 18*(5), 509-521.

Pylkkänen, L., & McElree, B. (2006). The syntax-semantics interface: On-line composition of sentence meaning. In M. J. Traxler & M. A. Gernsbacher (Eds.), *Handbook of psycholinguistics* (pp. 537–577). New York, NY: Elsevier.

Pylkkänen, L., & McElree, B. (2007). An MEG study of silent meaning. *Journal of cognitive neuroscience, 19*(11), 1905-1921. doi:10.1162/jocn.2007.19.11.1905

Rommers, J., Meyer, A. S., Praamstra, P., & Huettig, F. (2013). The contents of predictions in sentence comprehension: Activation of the shape of objects before they are referred to. *Neuropsychologia, 51*(3), 437-447.

Singer, W., & Gray, C. M. (1995). Visual feature integration and the temporal correlation hypothesis. *Annual review of neuroscience, 18*(1), 555-586.

Springer, K., & Murphy, G. L. (1992). Feature Availability in Conceptual Combination. *Psychological Science, 3*(2), 111-117.

Sternberg, S. (1969). Memory-scanning: Mental processes revealed by reaction-time experiments. *American scientist*, 421-457.

Treisman, A. (1996). The binding problem. *Current opinion in neurobiology, 6*(2), 171-178.

Von Der Malsburg, C. (1984). *The correlation theory of brain function* (1461287367). Retrieved from Göttingen, NL:

Werning, M., Hinzen, W., & Machery, E. (2012). *The Oxford handbook of compositionality*: Oxford University Press.

Zwaan, R. A., Stanfield, R. A., & Yaxley, R. H. (2002). Language comprehenders mentally represent the shapes of objects. *Psychological Science, 13*(2), 168-171.